

# NOTES 24: INFERENCE FOR REGRESSION

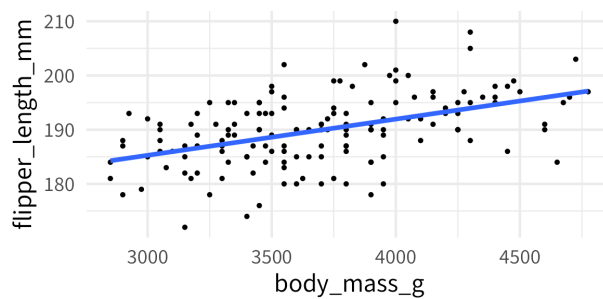
Stat 120 | Fall 2025

Prof Amanda Luby

---

When we fit linear regression models earlier in this class, we used them to describe the relationship between the variables and interpreted the slope and the intercept as descriptions of the data. Now, we'd like to understand what the regression model can tell us about the relationship of the variables *in the population*.

We're going to return to the `palmerpenguins` dataset from earlier in the class and restrict our analysis today to Adelie penguins. We want to predict `flipper_length_mm` using `body_mass_g`.



Call:

```
lm(formula = flipper_length_mm ~ body_mass_g, data = adelie)
```

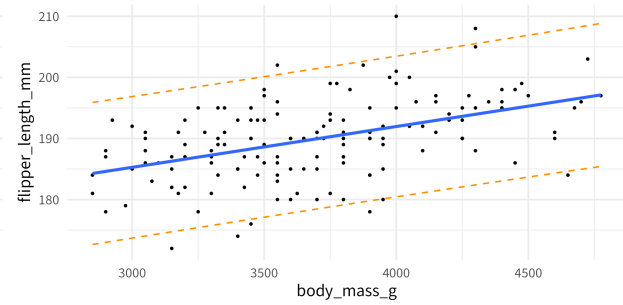
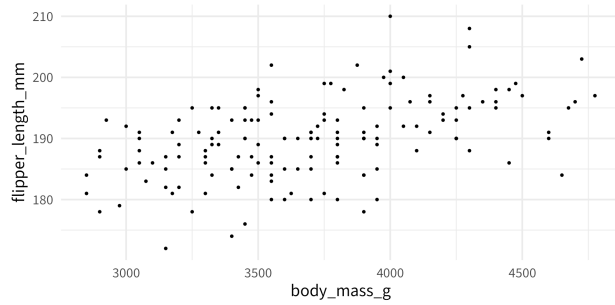
Coefficients:

(Intercept)	body_mass_g
1.65e+02	6.68e-03

**Warm up:** Write out the regression equation and provide an interpretation for the slope

The **population model** for the regression line is:

Idea:



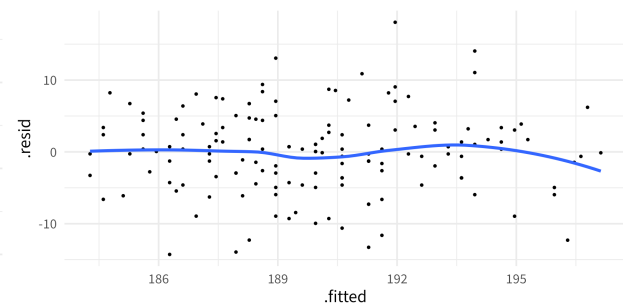
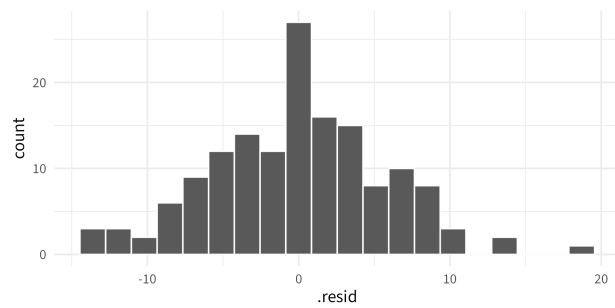
To check if this population model is reasonable, use the “LINE” mnemonic:

1. **L** \_\_\_\_\_

2. **I** \_\_\_\_\_

3. **N** \_\_\_\_\_

4. **E** \_\_\_\_\_



The **sampling distribution** for  $b_1$  is:

```
summary(adelie_lm)
```

Call:

```
lm(formula = flipper_length_mm ~ body_mass_g, data = adelie)
```

Residuals:

Min	1Q	Median	3Q	Max
-14.277	-3.619	0.057	3.470	18.048

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	1.65e+02	3.85e+00	42.93	< 2e-16 ***
body_mass_g	6.68e-03	1.03e-03	6.47	1.3e-09 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.8 on 149 degrees of freedom

(1 observation deleted due to missingness)

Multiple R-squared: 0.219, Adjusted R-squared: 0.214

F-statistic: 41.8 on 1 and 149 DF, p-value: 1.34e-09

**CLT-based confidence interval for  $\beta_1$ :**

$$b_1 \pm t_{n-2}^* \times SE_{b_1}$$

**CLT-based test for  $H_0 : \beta_1 = 0$ :**

$$t_{n-2} = \frac{b_1 - 0}{SE_{b_1}}$$

**Example:** Using the summary() output, along with the sampling distribution information above, how would you compute a 95% confidence interval for  $\beta_1$ ?

## 1 Inference for Predictions (Ch 9.3)

**Example:** We measure a new Adelie penguin that weighs 4000g. What is (a) the predicted *mean* flipper length among all 4000g penguins and (b) the predicted flipper length for *any* 4000g penguin?

```
augment(adelle_lm, newdata = tibble(body_mass_g = 4000))
```

```
# A tibble: 1 x 2
  body_mass_g .fitted
    <dbl>     <dbl>
1     4000     192.
```

```
augment(adelle_lm, newdata = tibble(body_mass_g = 4000), interval = "confidence")
```

```
# A tibble: 1 x 4
  body_mass_g .fitted .lower .upper
    <dbl>     <dbl> <dbl> <dbl>
1     4000     192.   191.   193.
```

```
augment(adelle_lm, newdata = tibble(body_mass_g = 4000), interval = "prediction")
```

```
# A tibble: 1 x 4
  body_mass_g .fitted .lower .upper
    <dbl>     <dbl> <dbl> <dbl>
1     4000     192.   180.   203.
```

We usually rely on R to make confidence and prediction intervals, but the formulas for these intervals are:

A \_\_\_\_\_ **interval for the mean response** when the predictor is  $x^*$  is:

$$\hat{y} \pm t_{n-2}^* \times s_e \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{(n-1)s_x^2}}$$

A \_\_\_\_\_ **interval for an individual response** when the predictor is  $x^*$  is:

$$\hat{y} \pm t_{n-2}^* \times s_e \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{(n-1)s_x^2}}$$